A class of smooth models satisfying marginal and context specific conditional independencies

R. Colombi[a], A. Forcina[b]

[a] *Statistical Laboratory, University of Bergamo, Italy*
[b] *Dipartimento di Economia, Finanza e Statistica, University of Perugia, Italy*

Abstract

We study a class of conditional independence models for discrete data with the property that one or more log-linear interactions are defined within two different marginal distributions and then constrained to 0; all the conditional independence models which are known to be non smooth belong to this class. We introduce a new marginal log-linear parameterization and show that smoothness may be restored by restricting one or more independence statements to hold conditionally to a restricted subset of the configurations of the conditioning variables. Our results are based on a specific reconstruction algorithm from log-linear parameters to probabilities and fixed point theory. Several examples are examined and a general rule for determining the implied conditional independence restrictions is outlined.

Keywords: categorical data, marginal log-linear parameterizations, smooth parameterizations.

1. Introduction

Conditional independence models for discrete data are determined by a set of constraints on log-linear interactions defined within different marginal distributions of a contingency table. The family of hierarchical and complete marginal log-linear parameterizations (HCMP for short) introduced by Bergsma and Rudas [4] provides a general framework for combining log-linear constraints defined on a collection of marginal distributions into an overall joint distribution. Methods for determining whether and how a conditional independence model may be translated into a HCMP have been studied by Rudas et al. [13] and Forcina et al. [9] among others; the fact that a HCMP exists, is a sufficient condition for the model to be smooth.

On the other hand, it is known that no HCMP exists when a model imposes constraints on the same log-linear interaction defined in two different marginals. It has been shown [4, Theorem 3] that, when the same interaction is defined in two different marginals, the jacobian of the mapping from log-linear parameters to probabilities is singular for the uniform distribution. Though, formally, this does not imply that the model itself has singularities, all known models with singularities correspond to cases where no HCMP exists because one or more interactions are constrained more than once. In this paper we study the class of conditional independence models where the same interaction is

constrained in two or more marginal distributions and we show, essentially, that any such model is non smooth but can be turned into a smooth model by restricting it to a suitable context specific conditional independence model.

Following Bergsma and Rudas [4], we may assume, without loss of generality, that the marginal distributions of interest have been arranged in a non decreasing order and that they will be reconstructed one at a time starting from the smallest. Because the full joint distribution is simply the last marginal in this list, we need only to consider how to determine a given marginal distribution when one or more log-linear interactions to be constrained have already been defined and/or constrained in a previous marginal. A useful tool for reconstructing marginal distributions in a sequence is the mixed parameterization [e.g., 2] by which we may combine the marginal probabilities from previous marginals with the log-linear interactions defined in the marginal distribution under consideration. Because the mapping produced by the mixed parameterization is one to one and smooth, the question of whether a model is smooth up to a given marginal, is equivalent to the question whether an algorithm based on the mixed parameterization exists and converges. By using results from the theory of fixed point algorithms, we study the jacobian of a new reconstruction algorithm that allows certain log-linear interactions to be redefined and show that this may either converge, and thus the model is smooth, remain at the starting point irrespective of the starting value, implying that the resulting distribution is not uniquely determined by the log-linear parameters or, simply not converge. A formal proof of these properties is derived under complete independence and we provide substantial evidence to support the conjecture that our results hold in the general case.

The results derived in this paper help clarifying which interaction parameters may be redefined and which other interactions should be omitted as a replacement. In particular we show that smoothness is restored only when a specific subset of other interactions is omitted; these interactions have the property that, when they are missing, and thus unconstrained, the conditional independence of interest holds only on a subset of the configuration of the conditioning variables. Log-linear models which allow context specific conditional independences have been studied in detail by Hojsgaard [10] who also derives a markov property for undirected graphs involving context specific conditional independencies. A special case of the results derived here was considered by Roverato et al. [12].

In section 2, we introduce the basic notations, define marginal log-linear interactions and review the properties of the mixed parameterization. In section 3, after presenting a set of motivating examples, we introduce a new algorithm for reconstructing a marginal distribution when interactions defined in previous marginals have to be constrained again and we analyze its convergence properties. In section 4 we study the consequences on the original conditional independence statements of omitting constraints on a specific subset of higher order interactions and show that this results in context specific restrictions.

2. Notations and preliminary results

We study the joint distribution of d discrete random variables where X_j, $j = 1, \ldots, d$, takes values in $(0, \ldots, r_j)$. For conciseness, we denote variables by their indices and

use capitals to denote non-empty subsets of $V = \{1, \ldots, d\}$; such subsets will determine the variables involved either in a marginal distribution or in an interaction term. The collection of all non-empty subsets of a set $M \subseteq V$ will be denoted by $\mathcal{P}(M)$. In the following we write $i_1 i_2 \ldots i_k$ as a shorthand notation for $\{i_1, i_2 \ldots, i_k\}$. For a given $M \subseteq V$, the marginal distribution in M is determined by the cell probabilities $p_M(\boldsymbol{x}_M) = P(X_j = x_j, \forall j \in M)$. We introduce a shorthand notation that allows to specify the values of selected subsets of the arguments in a marginal probability and on the log-linear interactions to be defined below. Let $J \subset I \subset M$, then $p_M(\boldsymbol{x}_J, \boldsymbol{x}_{\mathrm{I}\setminus\mathrm{J}}, \boldsymbol{x}_{\mathrm{M}\setminus\mathrm{I}})$ denotes the marginal probability where $\boldsymbol{x}_J$ is the value of X_h, $h \in J$, $\boldsymbol{x}_{\mathrm{I}\setminus\mathrm{J}}$ the value of X_h, $h \in \mathrm{I}\setminus\mathrm{J}$ and $\boldsymbol{x}_{\mathrm{M}\setminus\mathrm{I}}$ the values of X_h, $h \in \mathrm{M}\setminus\mathrm{I}$. We will also write $\mathbf{0}_{\mathrm{I}\setminus\mathrm{J}}$ to state that $X_h = 0$, $\forall h \in \mathrm{I}\setminus\mathrm{J}$.

2.1. Marginal and conditional log-linear interactions

Though there are many different ways of coding marginal log-linear parameters, parameters defined by different codings are linearly related; thus there is no loss of generality in using the *reference category* coding, where comparisons are with respect to the category taken as reference, usually the first.

Definition 1. *A reference category log-linear interaction I within M is defined by the following expression*

$$\eta_{I;M}(\boldsymbol{x}_I \mid \boldsymbol{x}_{M\setminus I}) = \sum_{J\subseteq I} (-1)^{|I\setminus J|} \log p_M(\boldsymbol{x}_J, \mathbf{0}_{I\setminus J}, \boldsymbol{x}_{M\setminus I}), \tag{1}$$

where, $\forall i \in I, x_i > 0$.

Example 1. *The logit of X_i at x_i computed within M is*

$$\eta_{i;M}(x_i \mid \boldsymbol{x}_{M\setminus i}) = \log p_M(x_i, \boldsymbol{x}_{M\setminus i}) - \log p_M(0_i, \boldsymbol{x}_{M\setminus i}), \quad x_i > 0$$

and the log-odds ratio for $X_i = x_i$, $X_j = x_j$ is

$$\begin{aligned} \eta_{H;M}(x_i, x_j \mid \boldsymbol{x}_{M\setminus H}) &= \log p_M(x_i, x_j, \boldsymbol{x}_{M\setminus H}) - \log p_M(0_i, x_j, \boldsymbol{x}_{M\setminus H}) \\ &- \log p_M(x_i, 0_j, \boldsymbol{x}_{M\setminus H}) + \log p_M(0_i, 0_j, \boldsymbol{x}_{M\setminus H}) \end{aligned}$$

where $H = i \cup j$.

It may be easily verified that, given $h \in \mathrm{M}\setminus\mathrm{I}$ and $H = I \cup h$, (1) implies the following recursive relation

$$\eta_{H;M}(\boldsymbol{x}_H \mid \boldsymbol{x}_{\mathrm{M}\setminus\mathrm{H}}) = \eta_{I;M}(\boldsymbol{x}_I \mid x_h, \boldsymbol{x}_{\mathrm{M}\setminus\mathrm{H}}) - \eta_{I;M}(\boldsymbol{x}_I \mid 0_h, \boldsymbol{x}_{\mathrm{M}\setminus\mathrm{H}}), \tag{2}$$

this indicates that interactions of higher order may be constructed by a sequence of first order differences starting from logits.

Whenever $\mathrm{M}\setminus\mathrm{I}$ is not empty, marginal log-linear interactions depend on the value of the remaining variables. Because (1) is a contrast of logarithms of marginal probabilities, it can be easily verified that $\eta_{I;M}(\boldsymbol{x}_I \mid \boldsymbol{x}_{\mathrm{M}\setminus\mathrm{I}})$ is the log-linear interaction I in the marginal

distribution M conditionally on $X_h = x_h \; \forall h \in \mathrm{M}\backslash\mathrm{I}$. Clearly, within the full collection of marginal log-linear interaction parameters conditional on the configurations of the remaining variables, there is a substantial amount of redundancy. Below we show that these parameters are linearly related and that they can all be written in terms of the subset where the conditioning variables are all fixed at their reference category; this subset contains non redundant elements.

For a given $\eta_{I;M}(\boldsymbol{x}_I \mid \boldsymbol{x}_{\mathrm{M}\backslash\mathrm{I}})$ let $h \in \mathrm{M}\backslash\mathrm{I}$, then (2) may be used to obtain

$$\eta_{I;M}(\boldsymbol{x}_I \mid \boldsymbol{x}_{\mathrm{M}\backslash\mathrm{I}}) = \eta_{I;M}(\boldsymbol{x}_I \mid 0_h, \boldsymbol{x}_{\mathrm{M}\backslash\mathrm{H}}) + \eta_{H;M}(\boldsymbol{x}_I, x_h \mid \boldsymbol{x}_{\mathrm{M}\backslash\mathrm{H}}).$$

Repeated use of the relation above leads to the following expansion

$$\eta_{I;M}(\boldsymbol{x}_I \mid \boldsymbol{x}_{\mathrm{M}\backslash\mathrm{I}}) = \sum_{I\subseteq H\subseteq M} \eta_{H;M}(\boldsymbol{x}_H \mid \boldsymbol{0}_{\mathrm{M}\backslash\mathrm{H}}). \tag{3}$$

The above equation shows that any marginal log-linear interaction may be written as a linear function of all possible higher order interactions conditional to the initial category of the remaining variables within the given marginal. For simplicity, in the following, we write $\eta_{I;M}(\boldsymbol{x}_I)$ as a shorthand for $\eta_{I;M}(\boldsymbol{x}_I \mid \boldsymbol{0}_{\mathrm{M}\backslash\mathrm{I}})$. An alternative way of removing conditioning variables, which has been applied to interactions defined as contrasts of averages of logarithms of probabilities, but could be applied to any type of interactions, is to average across the set of all possible configurations of the conditioning variables $\boldsymbol{x}_{\mathrm{M}\backslash\mathrm{I}}$. The log linear interactions used by Bergsma and Rudas [4], among others, are defined in this way; Lemma 8 in the Appendix shows that these interactions are linear functions of all the interactions $\eta_{H;M}(\boldsymbol{x}_J)$ for $H \supseteq I$.

Example 2. *Suppose that $M = I \cup h \cup k$, then*

$$\eta_{I;M}(\boldsymbol{x}_I \mid x_h, x_k) = \eta_{I;M}(\boldsymbol{x}_I) + \eta_{I\cup h;M}(\boldsymbol{x}_I, x_h) + \eta_{I\cup k;M}(\boldsymbol{x}_i, x_k) + \eta_{I\cup h\cup k;M}(\boldsymbol{x}_I, x_h, x_k).$$

For any $I \in \mathcal{P}(M)$, it is convenient to arrange the log-linear interactions $\eta_{I;M}(\boldsymbol{x}_I)$ into the vector $\boldsymbol{\eta}(I, M)$ with elements in lexicographic order of $\boldsymbol{x}_I$; this vector may be written as

$$\boldsymbol{\eta}(I, M) = \boldsymbol{C}(I, M) \log \boldsymbol{p}(M), \tag{4}$$

where $\boldsymbol{C}(I, M) = \bigotimes_{j=1}^{d} \boldsymbol{C}_j$ and $\boldsymbol{C}_j = \left(-\mathbf{1}_{r_j}\, \boldsymbol{I}_{r_j}\right)$ if $j \in I$ and $\boldsymbol{C}_j = \left(1, \mathbf{0}'_{r_j}\right)$ otherwise. Let also $\boldsymbol{\eta}(M) = \boldsymbol{C}(M) \log \boldsymbol{p}(M)$ denote the vector obtained by stacking the $\boldsymbol{\eta}(I, M)$ components one below the other in lexicographic order relative to $I \in \mathcal{P}(M)$. It is well known that under multinomial sampling, $\boldsymbol{\eta}(M)$ constitutes a vector of variation independent canonical parameters for $\boldsymbol{p}(M)$. Let $\boldsymbol{G}(I, M) = \bigotimes_{j=1}^{d} \boldsymbol{G}_j$, where $\boldsymbol{G}_j$ is an identity matrix of order $r_j + 1$ without the first columns if $j \in I$ and $\mathbf{1}_{r_j+1}$ otherwise. Let $\boldsymbol{G}(M)$ be the matrix whose columns are given by the $\boldsymbol{G}(I, M)$ matrices arranged one aside the other in lexicographic order. It is easily verified that $\boldsymbol{G}(M)$ is the right inverse of $\boldsymbol{C}(M)$; this implies the reconstruction formula

$$\log \boldsymbol{p}(M) = \boldsymbol{G}(M)\boldsymbol{\eta}(M) - \mathbf{1} \log\{\mathbf{1}' \exp[\boldsymbol{G}(M)\boldsymbol{\eta}(M)]\}. \tag{5}$$

2.2. The mixed parameterization

Within the distribution in M, the vector of *mean* parameters $\boldsymbol{\mu}_{\mathcal{P}(M)}$ [2, p. 121] is the expected value of the sufficient statistics for $\boldsymbol{\eta}(M)$ in a sample of size 1 and equals

$$\boldsymbol{\mu}_{\mathcal{P}(M)} = \boldsymbol{G}'(M)\boldsymbol{p}(M);$$

there is a diffeomorphism between $\boldsymbol{\mu}_{\mathcal{P}(M)}$ and $\boldsymbol{\eta}(M)$ [2, p. 121]. Because each block of rows $\boldsymbol{C}(I, M)$ in $\boldsymbol{C}(M)$ corresponds to a block of columns $\boldsymbol{G}(I, M)$ in $\boldsymbol{G}(M)$, we may define $\boldsymbol{\mu}(I) = G(I, M)'\boldsymbol{p}(M)$ to be the collection of mean parameters for a given interaction. It is worth noting that, though mean parameters, like canonical parameters, are associated to interactions $I \in \mathcal{P}(M)$, $\boldsymbol{\mu}(I)$ may be defined in any marginal such that $I \subseteq M$. Having coded the canonical parameters as contrasts with respect to the initial category, the corresponding mean parameters are simply marginal probabilities.

We recall a definition and a few results which are relevant in the following.

Definition 2. *For an arbitrary margin M, let $(\mathcal{U}, \mathcal{V})$ be a partition of $\mathcal{P}(M)$; the pair of vectors $[\boldsymbol{\eta}_{\mathcal{U},M}, \boldsymbol{\mu}_{\mathcal{V}}]$, where $\boldsymbol{\eta}_{\mathcal{U},M} = (\boldsymbol{\eta}(I, M), I \in \mathcal{U})$ is composed of canonical parameters, and $\boldsymbol{\mu}_{\mathcal{V}} = (\boldsymbol{\mu}(I), I \in \mathcal{V})$ is composed of mean parameters, constitute a* mixed parameterization *of the marginal distribution $\boldsymbol{p}(M)$.*

In the following, to be short, we will often refer to the log-linear parameters $\boldsymbol{\eta}_{\mathcal{U},M} = (\boldsymbol{\eta}(I, M), I \in \mathcal{U})$ as *log-linear interactions in $\mathcal{U}$* or *collection $\mathcal{U}$ of log-linear parameters.*

Lemma 1. *For any mixed parameterization, there is a diffeomorphism between the vector of mean parameters $\boldsymbol{\mu}_{\mathcal{P}(M)}$ and the pair of vectors $[\boldsymbol{\eta}_{\mathcal{U},M}, \boldsymbol{\mu}_{\mathcal{V}}]$; in addition, the two components are variation independent.*

Proof. See [2, p. 121-122] □

The numerical algorithm for reconstructing $\boldsymbol{p}(M)$ from $[\boldsymbol{\eta}_{\mathcal{U},M}, \boldsymbol{\mu}_{\mathcal{V}}]$ given by Forcina [8] is a faster alternative to the usual IPF algorithm.

The mixed parameterization is a powerful tool for reconstructing a joint distribution from marginal log-linear parameters because one can process one marginal distribution at a time by combining the log-linear parameters defined within that distributions with the mean parameters, or, equivalently, marginal probabilities, available from marginal distributions reconstructed in previous steps. As long as these two sets of interactions are a partition of $\mathcal{P}(M)$, the basic argument used by Bartolucci et al. [3] implies that any model defined by linear constraints on the marginal log-linear parameters constitutes a curved exponential family and thus is smooth.

3. The LM reconstruction algorithm

In this section we investigate the properties of conditional independence models which require to impose non trivial constraints on the same log-linear interactions defined in two or more marginal distributions. We may suppose, without loss of generality, that the marginals of interest are arranged in non decreasing order and that they will be processed

one at a time, starting from the first one. In this way, at each step in the reconstruction of the joint distribution from its marginal log-linear parameters, we need only be concerned with the marginal at hand and examine whether, by use of the mixed parameterization, we may combine the mean parameters from previous marginals with the log-linear parameters which are either available or need to be constrained in the marginal under consideration. An algorithm for doing this is presented and its convergence properties investigated.

3.1. Motivating examples

We now present a set of examples which will highlight different features of the kind of models we are going to consider. Each model is made of two parts: (i) a list of conditional independencies which have been accommodated, somehow, in previous marginals (ii) an additional conditional independence to be imposed in the current marginal M. We start with a couple of elementary models:

Example 3. *Suppose that, having assumed that* $1 \perp\!\!\!\perp 2 \mid 3$ *in the marginal 123, in the marginal* $M = 1234$ *we want also* $1 \perp\!\!\!\perp 2 \mid (3,4)$*. Here we need to constrain again the* $\{12, 123\}$ *interactions; in the binary case, Evans [7] has shown that the model has singularities.*

Example 4. *Suppose that, having assumed that* $1 \perp\!\!\!\perp (2,4)$ *in the marginal 124, we want also* $2 \perp\!\!\!\perp 4 \mid (1,3)$*. Here, in addition to the* 124 *interaction which has already been constrained in* 124*, we need to constrain* 24 *which was defined in the previous marginal; in the binary case, Drton [5] has shown that the model has singularities.*

The nest example is a little more complex:

Example 5. *Having assumed that* $1 \perp\!\!\!\perp 2 \mid 3$ *and* $1 \perp\!\!\!\perp 3 \mid 4$ *we also want* $1 \perp\!\!\!\perp (2,3) \mid (4,5)$*; here the list of interactions to be constrained again is given by* $\{12, 123, 13, 134\}$*.*

The following examples are different because the collection of interactions that have already been defined in previous marginal is too large to be redefined again in M:

Example 6. *Suppose that, having set* $1 \perp\!\!\!\perp 2 \mid (3,4)$ *and* $1 \perp\!\!\!\perp 2 \mid (3,5)$ *we also want* $1 \perp\!\!\!\perp 2 \mid (3,4,5)$*; here the collection of interactions that have already been defined and that have to be constrained again is* $\{12, 123, 124, 125, 1234, 1235\}$*.*

Example 7. *Suppose that, having set* $1 \perp\!\!\!\perp 2 \mid (3,4)$*,* $1 \perp\!\!\!\perp 2 \mid (3,5)$ *and* $1 \perp\!\!\!\perp 2 \mid (4,5)$ *we also want* $1 \perp\!\!\!\perp 2 \mid (3,4,5)$*, here all the interactions in the ascending class from* 12 *to* $M = 12345$*, except* M *itself, have to be constrained again.*

3.2. Setting up the framework

Let M denote the current marginal, $\mathcal{V}$ the collection of interactions defined in previous marginals which belong to $\mathcal{P}(M)$ and $\mathcal{L} = \mathcal{P}(M) \backslash \mathcal{V}$. Let also $\mathcal{A}$ be the collection of interactions to be constrained in M according to the last conditional independence statement; whenever $\mathcal{V} \cap \mathcal{A} \neq \emptyset$, we are trying to constrain again the corresponding log-linear interaction. Though we would like to redefine and constrain in M all the interactions in

$\mathcal{V} \cap \mathcal{A}$, we shall see that this is not always possible; denote by $\mathcal{I} \subseteq (\mathcal{V} \cap \mathcal{A})$ the actual collection which we redefine in M and $\mathcal{R} = \mathcal{V} \backslash \mathcal{I}$ the remaining interactions.

Because the mean parameters in $\mathcal{I} \cup \mathcal{R}$ together with the log-linear parameters in $\mathcal{L}$ constitute a mixed parameterization of $\boldsymbol{p}(M)$, these parameters determine uniquely the value of the log-linear parameters in $\mathcal{I}$ to be redefined within M; thus they cannot be constrained again, unless we remove from $\mathcal{L}$ a collection, say $\mathcal{H}$, of log-linear interactions with exactly the same number of parameters as the collection $\mathcal{I}$; below we investigate whether such an atypical parameterization may provide a smooth mapping. We shall see that the two sets $\mathcal{I}, \mathcal{H}$ must be chosen carefully and satisfy a set of conditions which establish a close relation between them.

Example 8. *Consider again example 6, here* $\mathcal{V} = \mathcal{P}(1234) \cup \mathcal{P}(1235)$, $\mathcal{A} = \{12, 123, 124, 125, 1234, 1235, 1245, 12345\}$ *and* $\mathcal{A} \cap \mathcal{V} = \{12, 123, 124, 125, 1234, 1235\}$*; as we shall see, not all the elements of this collection can be redefined in* M*, the most we can achieve is to set* $\mathcal{I} = \{12, 123\}$ *and* $\mathcal{H} = \{1245, 12345\}$ *where* X_4 *and* X_5 *are fixed to a given category.*

Example 9. *In example 4,* $\mathcal{V} = \mathcal{P}(124)$, $\mathcal{L} = \mathcal{P}(1234) \backslash \mathcal{P}(124)$*; suppose we set* $\mathcal{I} = \{24, 124\}$ *and* $\mathcal{H} = \{234, 1234\}$ *where* X_3 *is fixed to a given category; it can be easily checked that* $\mathcal{H}$ *indexes the same number of parameters as* $\mathcal{I}$*.*

3.3. Description of the algorithm

The problem, when reconstructing the distribution in M, is how to combine the mean parameter $\boldsymbol{\mu}_{\mathcal{I}}$, available from previous marginals with the log-linear parameters $\boldsymbol{\eta}_{\mathcal{I};M}$ defined again in the present marginal. Recall that the mixed parameterization require that mean parameters and log-linear interactions must refer to two complementary sets whose union is $\mathcal{P}(M)$. The idea is to remove from the log-linear parameters $\mathcal{L}$, to be defined in M, the subset $\mathcal{H}$ with the same number of parameters as the elements of $\mathcal{I}$. The algorithm that we describe below can handle such a context and the issue will be to determine under which conditions such an algorithm may converge; if it does, then it can be shown that the model is smooth. The algorithm for reconstructing the marginal distribution in M is made of two steps and require starting values for $\boldsymbol{\eta}_{\mathcal{H};M}$:

M-step given the latest guess for the log-linear parameters $\boldsymbol{\eta}_{\mathcal{H};M}$, an updated estimate for the vector of mean parameters $\boldsymbol{\mu}_{\mathcal{H}}$ may be computed by a mixed parameterization with mean parameters indexed by the collection of interactions $\mathcal{R} \cup \mathcal{I}$ and log-linear parameters indexed by $\mathcal{L}$;

L-step given the latest guess for the vector of mean parameters $\boldsymbol{\mu}_{\mathcal{H}}$, an updated estimate for the vector of log-linear parameters $\boldsymbol{\eta}_{\mathcal{H};M}$ may be computed by a mixed parameterization with mean parameters indexed by $\mathcal{R} \cup \mathcal{H}$ and log-linear parameters indexed by $\mathcal{I} \cup (\mathcal{L} \backslash H)$.

In order to examine the properties of the LM algorithm, we need to determine how changes in the input value of $\boldsymbol{\eta}_{\mathcal{H};M}$ in the M step affects the output value produced in the L step. For this purpose, we recall results concerning the derivatives of certain components of the mixed parameterization relative to others which are relevant here. In the following

write $\boldsymbol{\pi}$ as a shorthand for $\boldsymbol{p}(M)$, let $\boldsymbol{D}_{\boldsymbol{\pi}} = \text{diag}(\boldsymbol{\pi})$ and let $\boldsymbol{\Omega} = \boldsymbol{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}\boldsymbol{\pi}'$ denote the derivative of $\boldsymbol{\pi}$ with respect to $\boldsymbol{\eta}(M)'$.

Lemma 2.

$$\boldsymbol{F}(M) = \frac{\partial \boldsymbol{\mu}_{\mathcal{P}(M)}}{\partial \boldsymbol{\eta}(M)'} = \boldsymbol{G}(M)'\boldsymbol{\Omega}(M)\boldsymbol{G}(M),$$

is the covariance matrix of a collection of distinct binary variables determined by the columns of $\boldsymbol{G}(M)$ *and thus is positive definite.*

Proof. See Forcina [8]. □

Any two subsets of interactions $\mathcal{H}, \mathcal{K} \subseteq \mathcal{P}(M)$ determine two sub-collections of binary random variables and a block in the covariance matrix $\boldsymbol{F}(M)$. In the following we omit reference to the marginal M when it is obvious from the context and write

$$\boldsymbol{F}_{\mathcal{H}\mathcal{K}} = \boldsymbol{G}'_{\mathcal{H}}\boldsymbol{\Omega}\boldsymbol{G}_{\mathcal{K}}.$$

Lemma 3. *In the M-step, where* $\mathcal{H}$ *is part of the log-linear parameter*

$$\frac{\partial \boldsymbol{\mu}_{\mathcal{H}}}{\partial \boldsymbol{\eta}'_{\mathcal{H};M}} = \boldsymbol{B} = \boldsymbol{F}_{\mathcal{H}\mathcal{H}} - \boldsymbol{F}_{\mathcal{H}\mathcal{V}}\boldsymbol{F}_{\mathcal{V}\mathcal{V}}^{-1}\boldsymbol{F}_{\mathcal{V}\mathcal{H}};$$

in the L-step, where $\mathcal{H}$ *is part of the mean parameter*

$$\frac{\partial \boldsymbol{\eta}_{\mathcal{H};M}}{\partial \boldsymbol{\mu}_{\mathcal{H}}} = \boldsymbol{A}^{-1} = \left(\boldsymbol{F}^{-1}_{(\mathcal{R}\cup\mathcal{H})(\mathcal{R}\cup\mathcal{H})}\right)_{\mathcal{H}\mathcal{H}} = \left(\boldsymbol{F}_{\mathcal{H}\mathcal{H}} - \boldsymbol{F}_{\mathcal{H}\mathcal{R}}\boldsymbol{F}^{-1}_{\mathcal{R}\mathcal{R}}\boldsymbol{F}_{\mathcal{R}\mathcal{H}}\right)^{-1},$$

where we have used the formula for the inverse of a partitioned matrix.

Proof. the result follows from Lemma 4 in Forcina [8]. □

A full step of the LM algorithm may be seen as a fixed point function which, given a guess value of $\boldsymbol{\eta}_{\mathcal{H};M}$, produces an updated estimate of the same vector. A sufficient condition for an algorithm to be a contraction [see for example 1], a property which implies that it converges to a unique solution, is that the jacobian of a full LM step has spectral radius (maximum absolute eigenvalue) strictly smaller than 1. Let $\boldsymbol{J} = \boldsymbol{A}^{-1}\boldsymbol{B}$ be the jacobian of this mapping; let also $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}} = \boldsymbol{F}_{\mathcal{I}\mathcal{H}} - \boldsymbol{F}_{\mathcal{I}\mathcal{R}}\boldsymbol{F}^{-1}_{\mathcal{R}\mathcal{R}}\boldsymbol{F}_{\mathcal{R}\mathcal{H}}$. An upper bound for the spectral radius of $\boldsymbol{J}$ is determined in the following lemma.

Lemma 4. *The spectral radius of* $\boldsymbol{J}$ *is always less than 1 except when* $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$ *is not of full rank.*

Proof. See the Appendix. □

The main result of this section is contained in the following Theorem and concerns the properties of the mapping from $\boldsymbol{\xi} = (\boldsymbol{\eta}_{\mathcal{L}\cup\mathcal{I}\setminus\mathcal{H},M}, \boldsymbol{\mu}_{\mathcal{V}})$ to $\boldsymbol{\pi}$, under the assumption that the elements of $\boldsymbol{\xi}$ are compatible, that is there is at least a $\boldsymbol{\pi}$ with the parameters specified by $\boldsymbol{\xi}$. The result depends on the spectral radius of the jacobian matrix $\boldsymbol{J}$ defined above.

Theorem 1. *Under the assumption that the elements of $\boldsymbol{\xi}$ are compatible, when $\boldsymbol{Q}_{\mathcal{IH}|\mathcal{R}}$ is of full rank, the mapping from $(\boldsymbol{\eta}_{\mathcal{L}\cup\mathcal{I}\setminus\mathcal{H},M}, \boldsymbol{\mu}_{\mathcal{V}})$ to $\boldsymbol{\pi}$ is one to one and smooth. In the special case when $\boldsymbol{Q}_{\mathcal{IH}|\mathcal{R}} = \mathbf{0}$, so that $\boldsymbol{J}$ is an identity matrix, the mapping is not one to one. When $\boldsymbol{Q}_{\mathcal{IH}|\mathcal{R}}$ is singular but different from a null matrix, the algorithm does not converge and nothing can be said about the smoothness of the mapping.*

Proof. Consider the sequence of vectors produced by the LM algorithm: $\boldsymbol{\eta}^{(0)}_{\mathcal{H};M}, \boldsymbol{\eta}^{(1)}_{\mathcal{H};M}, \ldots,$ where $\boldsymbol{\eta}^{(0)}_{\mathcal{H};M}$ is the starting value and $\boldsymbol{\eta}^{(s)}_{\mathcal{H};M}$ is the output of one step of the LM algorithm when we use $\boldsymbol{\eta}^{(s-1)}_{\mathcal{H};M}$ as input; because we have assumed that there is at least a compatible solution inside the parameter space, [1, Theorem 1.1] implies that, if the spectral radius of $\boldsymbol{J}$ is strictly less than one, the sequence converges to a unique solution. At convergence the argument in [3, Theorem 1] can be applied to show that the mapping is a diffeomorphism. In the special case when the jacobian matrix $\boldsymbol{J}$ is an identity matrix, $\boldsymbol{\eta}^{(0)}_{\mathcal{H};M} = \boldsymbol{\eta}^{(1)}_{\mathcal{H};M}$, so the algorithm converges in one step, irrespective of the starting value. This implies that, if $\boldsymbol{\pi}^{(0)}$ is the probability vector corresponding to $\boldsymbol{\eta}^{(0)}_{\mathcal{H};M}$ there is a whole neighbourhood of $\boldsymbol{\pi}^{(0)}$ whose points share exactly the same vector $\boldsymbol{\xi}$ of mean and log-linear parameters. □

Remark 1. *According to Theorem 1, a model may be smooth even if the log-linear interactions in $\mathcal{I}$ are defined and constrained in two different marginals. This is apparently in conflict with the result of [4, Theorem 3] which says that the jacobian obtained by differentiating the same log-linear interaction I defined in two different marginals, say M_1, M_2, with respect to $\boldsymbol{p}$, is singular for the uniform distribution, a condition which is necessary (but not sufficient) for a model to have singularities. However, when the set $M\setminus I$ is not empty, the log-linear interactions defined by Bergsma and Rudas [4] are constructed by averaging conditional interactions across all possible configurations of the conditioning variables. As mentioned in section 2.1, the results of Lemma 8 in the Appendix imply that any constraint on one of their log-linear interactions is equivalent to a linear constraint on the whole ascending class of our interactions with minimal element I and maximal element M. Hence the LM algorithm is not directly applicable to interactions defined in that way.*

3.4. Convergence of the algorithm

Below we derive a more convenient expression for $\boldsymbol{Q}_{\mathcal{IH}|\mathcal{R}}$ and show that the matrix is non singular under complete independence, if the set $\mathcal{H}$ satisfies certain conditions. We also determine conditions under which $\boldsymbol{Q}_{\mathcal{IH}|\mathcal{R}}$ is singular or null. Finally, we discuss the singularity of the same matrix when $\boldsymbol{p}(M)$ is unrestricted.

Let $\boldsymbol{P}_\emptyset = \mathbf{1}\boldsymbol{\pi}'$ be the projector, according to the metric defined by the matrix $\boldsymbol{D}_{\boldsymbol{\pi}}$, on the space spanned by the vector $\mathbf{1}$. By simple algebra, it can be shown that:

$$\boldsymbol{F}_{\mathcal{IH}} \;=\; \boldsymbol{G}'_{\mathcal{I}}\boldsymbol{\Omega}\boldsymbol{G}_{\mathcal{H}} = \boldsymbol{G}'_{\mathcal{I}}(\boldsymbol{I}-\boldsymbol{P}_\emptyset)'\boldsymbol{D}_{\boldsymbol{\pi}}(\boldsymbol{I}-\boldsymbol{P}_\emptyset)\boldsymbol{G}_{\mathcal{H}}.$$

From the previous result, it follows that:

$$\boldsymbol{Q}_{\mathcal{IH}|\mathcal{R}} = \boldsymbol{G}'_{\mathcal{I}}\boldsymbol{D}_{\boldsymbol{\pi}}(\boldsymbol{I}-\boldsymbol{P}_{\overline{R}})(\boldsymbol{I}-\boldsymbol{P}_\emptyset)\boldsymbol{G}_{\mathcal{H}},$$

where

$$\boldsymbol{P}_{\overline{R}} = (\boldsymbol{I} - \boldsymbol{P}_{\emptyset})\boldsymbol{G}_{\mathcal{R}}\boldsymbol{F}_{\mathcal{R}\mathcal{R}}^{-1}\boldsymbol{G}_{\mathcal{R}}'(\boldsymbol{I} - \boldsymbol{P}_{\emptyset})'\boldsymbol{D}_{\boldsymbol{\pi}}$$

is the projector, according to the metric defined by the matrix $\boldsymbol{D}_{\boldsymbol{\pi}}$, on the space spanned by the columns of $(\boldsymbol{I} - \boldsymbol{P}_{\emptyset})\boldsymbol{G}_{\mathcal{R}}$.

Let $\mathcal{S}(\boldsymbol{X})$ denote the space spanned by the columns of $\boldsymbol{X}$, for every $a \subseteq M$ let also $\boldsymbol{X}_a = \bigotimes_{j \in M} \boldsymbol{X}_j$, where $\boldsymbol{X}_j = \boldsymbol{I}_j$ if $j \in a$ and $\boldsymbol{X}_j = \boldsymbol{1}_j$ otherwise, and let $\boldsymbol{P}_a = \boldsymbol{X}_a(\boldsymbol{X}_a'\boldsymbol{D}_{\boldsymbol{\pi}}\boldsymbol{X}_a)^{-1}\boldsymbol{X}_a'\boldsymbol{D}_{\boldsymbol{\pi}}$ be the projection matrix onto $\mathcal{S}(\boldsymbol{X}_a)$. Let $\boldsymbol{X}_{\mathcal{I}\cup\mathcal{R}}$ be the matrix made by the columns of $\boldsymbol{X}_a \forall a \in \mathcal{I}\cup\mathcal{R}$ and $\boldsymbol{P}_{\mathcal{I}\cup\mathcal{R}}$ the projection onto $\mathcal{S}(\boldsymbol{X}_{\mathcal{I}\cup\mathcal{R}})$. Because $\mathcal{S}(\boldsymbol{G}_{\mathcal{R}})$ and $\mathcal{S}(\boldsymbol{1})$ belong to $\mathcal{S}(\boldsymbol{X}_{\mathcal{I}\cup\mathcal{R}})$ the projection matrix $\boldsymbol{P}_{\mathcal{I}\cup\mathcal{R}}$ commutes with both $\boldsymbol{P}_{\overline{R}}$ and $\boldsymbol{P}_{\emptyset}$; in addition, by using the identity $\boldsymbol{P}_{\mathcal{I}\cup\mathcal{R}}\boldsymbol{G}_{\mathcal{I}} = \boldsymbol{G}_{\mathcal{I}}$ it follows that

$$\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}} = \boldsymbol{G}_{\mathcal{I}}'\boldsymbol{D}_{\boldsymbol{\pi}}(\boldsymbol{I} - \boldsymbol{P}_{\overline{R}})(\boldsymbol{I} - \boldsymbol{P}_{\emptyset})\boldsymbol{P}_{\mathcal{I}\cup\mathcal{R}}\boldsymbol{G}_{\mathcal{H}}. \tag{6}$$

3.4.1. The case of complete independence

Lemma 5. *Under complete independence of the variables in M, (i) if $\mathcal{H}$ contains an interaction $v \notin \mathcal{A}$, the corresponding columns in $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$ are null, (ii) if $\mathcal{H}$ contains an interaction v where at least one of the variables in v is not binary and not contained in any element of $\mathcal{V}$, $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$ has a block of columns which is not of full rank.*

Proof. See the Appendix □

Lemma 5 suggests two necessary conditions for $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$ to be non singular: $\mathcal{H}$ cannot contain interactions not in $\mathcal{A}$ and all the non binary variables involved in the class of interactions $\mathcal{H}$ and not present in the class $\mathcal{I}\cup\mathcal{R}$, must be fixed to a single category different from the reference category. This implies that only a limited number of higher order interactions in M can be used as a replacement for those in $\mathcal{I}$. The definition below provides a set of conditions for $\mathcal{H}$ which will be shown to be sufficient. Let $\mathcal{K} = \{m_1, \ldots, m_r\}$ be the family of the maximal sets of $\mathcal{I}\cup\mathcal{R}$; for $t \in \mathcal{I}$, let $K(t) = \{m : m \in \mathcal{K}, t \subseteq m\}$ be the family of sets $m \in \mathcal{K}$ that contain t, $\mathcal{K}(t,h)$ be the family of the sets $\mathcal{G}$, $\mathcal{G} \in \mathcal{P}[\mathcal{K}(t)]$, such that $h \cap \bigcap_{m_j \in \mathcal{G}} m_j = \emptyset$ and $\bar{\mathcal{K}}(t,h) = \mathcal{P}(\mathcal{K})\backslash\mathcal{K}(t,h)$.

Definition 3. *A set $\mathcal{H}$ is a valid replacement for a given $\mathcal{I}$ if it satisfies the following conditions:*

(i) there is a one to one correspondence between the elements of $\mathcal{I}$ and $\mathcal{H}$ such that, for each $t \in \mathcal{I}$, there is a $v = t \cup h \in \mathcal{H}$, $t \cap h = \emptyset$, where the variables in h are fixed to a given category different from the reference category;

(ii) $\sum_{\mathcal{G}\in\mathcal{K}(t,h)} (-1)^{|\mathcal{G}|} \neq 0$;

(iii) there exists a complete ordering "$\prec$" in $\mathcal{I}$, coherent with the partial ordering of set inclusion, such that, for every $t \cup h \in \mathcal{H}$, $\mathcal{G} \in \bar{\mathcal{K}}(t,h)$ and $s = \left(\bigcap_{m_j\in\mathcal{G}} m_j\right) \cap (t \cup h)$ either $s \in \mathcal{R}$, or $s \in \mathcal{I}$ and $s \prec t$.

To clarify these notions, we discuss a few examples where we write (t,h) as a shorthand for $t \cup h$ if $t \cup h \in \mathcal{H}$.

Example 10. *In example 3 with* $\mathcal{I} = \{12, 123\}$ *and* $\mathcal{H} = \{(12, 4), (123, 4)\}$*, all conditions are trivially satisfied with* $\mathcal{K} = \{123\}$*; this is also the only element of* $\mathcal{K}(t, h)$*,* $\forall t, h$*, the same happens in example 4. In example 5 with the ordered set* $\mathcal{I} = \{12, 13, 123, 134\}$ *and* $\mathcal{H} = \{(12, 5), (13, 5), (123, 5), (134, 5)\}$ *condition (i) is clearly satisfied. In this case we have:* $\mathcal{K} = \{123, 134\}$*,* $\mathcal{K}(134, 5) = \{134\}$*,* $\mathcal{K}(123, 5) = \{123\}$*,* $\mathcal{K}(13, 5) = \{123, 134, \{123, 134\}\}$ *and* $\mathcal{K}(12, 5) = \{123\}$ *and condition (ii) is satisfied by these sets. For* $\bar{\mathcal{K}}(134, 5) = \{123, \{123, 134\}\}$ *condition (iii) is satisfied with* $s = 13$*. The set* $\bar{\mathcal{K}}(123, 5) = \{134, \{123, 134\}\}$ *satisfies (iii) because s=13. In the case of* $\bar{\mathcal{K}}(12, 5) = \{134, \{123, 134\}\}$ *(iii) holds because* $s = 1$*. The family* $\bar{\mathcal{K}}(13, 5)$ *is empty and so in this case condition (iii) is void.*

Example 11. *Having assumed* $1 \perp\!\!\!\perp 2 \mid (3, 4)$*,* $1 \perp\!\!\!\perp 2 \mid (3, 5)$*,* $1 \perp\!\!\!\perp 2 \mid (3, 6)$*, we also want* $1 \perp\!\!\!\perp 2 \mid (4, 5, 6)$*. It can be verified that* $\mathcal{H} = \{(12, 56), (124, 56)\}$ *is a valid replacement for* $\mathcal{I} = \{12, 124\}$*; here* $\mathcal{K} = \{124, 125, 126\}$ *and* $\mathcal{K}(124, 56) = \{124\}$*. It is easy to see that (i) and (ii) are satisfied. Condition (iii) holds in the case of* $\bar{\mathcal{K}}(124, 56) = \{125, 126, \{124, 125\}, \{124, 126\}, \{125, 126\}, \{124, 125, 126\}\}$*, because apart from the first two elements which produce sets s that belong to* $\mathcal{R}$*, all the others produce* $s = 12$*. A similar remark holds in that case of* $\mathcal{K}(124, 56)$*. However, if we set* $\mathcal{I} = \{12, 124, 125, 126\}$*, though* $\mathcal{H} = \{(12, 56), (124, 56), (125, 4), (126, 4)\}$ *satisfies conditions (i) and (ii), (iii) does not hold.*

We now give an instance where condition (ii) is not satisfied.

Example 12. *Suppose that* $1 \perp\!\!\!\perp 2 \mid (3, 4)$*,* $1 \perp\!\!\!\perp 2 \mid (3, 5)$ *and finally* $1 \perp\!\!\!\perp 2 \mid (3, 4, 5, 6)$*. Though the best choice would be to set* $\mathcal{I} = \{12, 123, 124, 125, 1234, 1235\}$*, if we set* $\mathcal{I} = \{12, 123\}$ *and* $\mathcal{H} = \{(12, 46), (123, 46)\}$*, condition (ii) is not satisfied.*

Lemma 6. *A pair* $\mathcal{I}$*,* $\mathcal{H}$*, where* $\mathcal{H}$ *is a valid replacement, always exists; for instance, take* $\mathcal{I} = \{t\}$*, where t is one of the minimal elements of* $\mathcal{A}$*, and* $\mathcal{H} = \{t \cup h\}$*, where h contains all the variables that belong to at most one element of* $\mathcal{K}(t)$ *when* $\mathcal{K}(t)$ *is not a singleton and by the variables that do not belong to the unique element of* $\mathcal{K}(t)$ *otherwise.*

Proof. See the Appendix □

Example 13. *In example 5, the minimal element t of* $\mathcal{A}$ *can be 12 or 13. If* $t = 12$ *then* $\mathcal{K}(t) = \{123\}$ *is a singleton. In this case h=45 and* $\mathcal{H}$ *contains only* $(12, 45)$*, the family* $\mathcal{K}(t, h)$ *contains only* $\{123\}$*. Instead, if we set* $t = 13$*,* $\mathcal{K}(t) = \{123, 134\}$ *and we must set* $h = 45$*, thus* $\mathcal{K}(t, h)$ *contains only the set* $\mathcal{G} = \{123, 134\}$*. In example 12,* $\mathcal{K} = \{1234, 1235\}$*, with* $t = 12$ *we must set* $h = 456$ *and* $\{1234, 1235\}$ *is the only set in* $K(t, h)$*. In example 11 with* $t = 12$*,* $\mathcal{K}(t) = \{1234, 1235, 1236\}$*, thus we must set* $h = 456$*; here* $\mathcal{K}(t, h)$ *has 3 elements of size 2 and 1 element of size 3 and the sum in condition (ii) of Definition 1 is -2.*

Let $\boldsymbol{G}_{t,h}(\boldsymbol{j}_h)$ be the sub-matrix of $\boldsymbol{G}_{t\cup h}$ where variables in h are fixed to $\boldsymbol{j}_h$.

Lemma 7. *Under complete independence:*
a) $\mathcal{S}(\boldsymbol{P}_a \boldsymbol{G}_{t,h}(\boldsymbol{j}_h)) \subseteq \mathcal{S}(G_r)$*, where* $r = a \cap (t \cup h)$*,*
b) if $t \subseteq a$ *and* $a \cap h = \emptyset$*,* $\boldsymbol{P}_a \boldsymbol{G}_{t,h}(\boldsymbol{j}_h) = \boldsymbol{G}_t P(\boldsymbol{x}_h = \boldsymbol{j}_h)$*.*

Proof. See the Appendix. □

Theorem 2. *If $\mathcal{H}$ is a valid replacement for $\mathcal{I}$, under complete independence, $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$ is non singular.*

Proof. Because under complete independence the projectors $\boldsymbol{P}_a$, $a \subseteq M$ commute, we can write $\boldsymbol{P}_{\mathcal{I}\cup\mathcal{R}} = \sum_{\mathcal{G}\in\mathcal{P}(\mathcal{K})}(-1)^{1+|\mathcal{G}|}\prod_{m\in\mathcal{G}}\boldsymbol{P}_m$, it follows that:

$$\boldsymbol{P}_{\mathcal{I}\cup\mathcal{R}}\boldsymbol{G}_{t,h}(\boldsymbol{j}_h) = \sum_{\mathcal{G}\in\mathcal{K}(t,h)}(-1)^{1+|\mathcal{G}|}\prod_{m\in\mathcal{G}}\boldsymbol{P}_m\boldsymbol{G}_{t,h}(\boldsymbol{j}_h) + \sum_{\mathcal{G}\in\bar{\mathcal{K}}(t,h)}(-1)^{1+|\mathcal{G}|}\prod_{m\in\mathcal{G}}\boldsymbol{P}_m\boldsymbol{G}_{t,h}(\boldsymbol{j}_h).$$

Condition (ii) of definition 3 and b) in lemma 7 imply that the first sum is $k\boldsymbol{G}_t$, with $k \neq 0$. Condition (iii) of definition 3 and a) in lemma 7 imply that, when an element in the second sum, say $\boldsymbol{U}$, is such that $\mathcal{S}(\boldsymbol{U}) \subseteq \mathcal{S}(\boldsymbol{G}_{\mathcal{R}})$, when we left multiply by $(\boldsymbol{I}-\boldsymbol{P}_{\bar{\mathcal{R}}})(\boldsymbol{I}-\boldsymbol{P}_{\emptyset})$, we get a null matrix because $(\boldsymbol{I}-\boldsymbol{P}_{\bar{\mathcal{R}}})$ projects onto the space orthogonal to $(\boldsymbol{I}-\boldsymbol{P}_{\emptyset})\boldsymbol{G}_{\mathcal{R}}$. In all other cases there exists a non null matrix $\boldsymbol{A}_{s,t}$, $s \prec t$, such that:

$$(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{R}})(-1)^{1+|\mathcal{G}|}\prod_{m\in\mathcal{G}}\boldsymbol{P}_m\boldsymbol{G}_{t,h}(\boldsymbol{j}_h) = (\boldsymbol{I}-\boldsymbol{P}_{\mathcal{R}})\boldsymbol{G}_s\boldsymbol{A}_{s,t}.$$

The matrix $\boldsymbol{G}_{\mathcal{H}}$ is made of blocks of columns of the form $\boldsymbol{G}_{t,h}(\boldsymbol{j}_h)$ and we may assume, without loss of generality, that these blocks are in the same order as the elements of $\mathcal{I}$ specified in condition (iii), then it follows that

$$\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}} = \boldsymbol{Q}_{\mathcal{I}\mathcal{I}|\mathcal{R}}\boldsymbol{A},$$

where the matrix $\boldsymbol{A}$ has blocks $\boldsymbol{A}_{s,t}$, $\quad s,t \in \mathcal{I}$ such that $\boldsymbol{A}_{s,t} = \boldsymbol{0}$ if $t \prec s$ and, because of condition (ii) of definition 3, the diagonal blocks $\boldsymbol{A}_{t,t}$, are proportional to an identity matrix, thus $\boldsymbol{A}$ is lower triangular and non singular. The result follows because both matrices in the product above are non singular. □

Example 14. *The choice of $\mathcal{I}$, $\mathcal{H}$ in the second part of example 11 does not satisfies (ii) still, numerical simulations indicate that $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$ is non singular. This exemplifies that the conditions of being adequate for replacement are only sufficient.*

Remark 2. *Theorem 6 of Roverato et al. [12] implies that, in the binary case, the model defined by $a \perp\!\!\!\perp b$ and $a \perp\!\!\!\perp b \mid c$, with all the elements of c equal 0, is smooth. If we set $\mathcal{I} = \mathcal{P}(a \cup b) \setminus (\mathcal{P}(a) \cup \mathcal{P}(b))$ and $\mathcal{H} = \{t, h : t \in \mathcal{I},\ h = c(\boldsymbol{j}_c)\}$ with $\boldsymbol{j}_c = \boldsymbol{1}$, our Theorem 2 implies that, under independence, the model $a \perp\!\!\!\perp b$ and $a \perp\!\!\!\perp b \mid c$, except when all the elements of c are equal 1, is smooth.*

3.4.2. The general case

Unfortunately, the main arguments used above depend crucially on the assumption of complete independence. For discrete data, all the models which have been shown to be non smooth, have a singular locus which is a subset of that defined by complete independence; instead, for gaussian models [6, Example 4.4] indicates that a non smooth

model may have a singular locus not contained in the model of complete independence. It is also interesting to note that numerical evaluations of $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$ outside the space of complete independence, indicate that it is of full rank even if $\mathcal{H}$ does not satisfies the conditions of Definition 3.

Taking into account all of the above, the extensive simulations which we have performed seem to support the conjecture that all the models obtained by replacing the interactions in $\mathcal{I}$ with an adequate replacement in $\mathcal{H}$ are indeed smooth everywhere in the parameter space. Though very unlikely, we cannot rule out the possibility that the models above may be singular on points of the parameter space which are outside the subspace defined by the model of complete independence. However, the expression for $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$ is very easy to compute and the software in MatLab and R which we provide as supplementary material, may be used for quick numerical checks.

4. Context specific conditional independence models

We have seen that a non smooth model may be transformed into a smooth one by omitting a class of log-linear interactions $\mathcal{H}$ to make place for other interactions defined in a previous marginal which we want to redefine in M. This implies that the values of the interactions in $\mathcal{H}$ are uniquely determined by the probabilities reconstructed in previous marginals and the log-linear interactions defined in M, thus they cannot be constrained. Because $\mathcal{H} \subseteq \mathcal{A}$, certain constraints implied by the original conditional independence in M cannot be implemented; in addition, when $\mathcal{A} \cap \mathcal{R} \neq \emptyset$, further limitations must be taken into account. In this section we study the nature and scope of the actual constraints that can be imposed in M. We remind that a conditional independence statement, that holds only on a subset of the configurations of the conditioning variables, is a context specific conditional independence (Hojsgaard [10]).

The collection of interactions $\mathcal{J} = \mathcal{H} \cup (\mathcal{A} \cap \mathcal{R})$ belongs to $\mathcal{A}$ but cannot be constrained in M; these interactions are of two kinds: (i) those which belong to $\mathcal{H}$ have to be omitted as a replacement for duplicating those in $\mathcal{I}$ and (ii) those in $\mathcal{R}$ which we were unable to replicate because, if included in $\mathcal{I}$, there would not exist an $\mathcal{H}$ adequate for replacement. It follows that the conditional independence in M must be restricted to the context that does not require to constrain the collection of interactions $\mathcal{J}$. More precisely, point (i) implies that the conditional independence can be defined only in the contexts in which, for every $(t, h) \in \mathcal{H}$ the variables in h are different from $\boldsymbol{j}_h$. Point (ii) implies that the conditional independence can be defined only in the contexts where, for every maximal set m of $\mathcal{I}$ and $v \in \mathcal{A} \cap \mathcal{R}$ the variables belonging to $v \setminus m$ are fixed to the reference category. Let us examine some of the previous example to clarify the situation.

Example 15. *Consider again example 5 here $\mathcal{A} \cap \mathcal{R}$ is empty and*

$$\mathcal{H} = \{(12, 5), (123, 5), (13, 5), (134, 5)\};$$

it follows that we are left with $1 \perp\!\!\!\perp (2, 3) \mid (4, 5)$, for all $X_5 \neq j_5$. In example 6, because $\mathcal{J} = \{124, 125, 1234, 1235, (12, 45), (123, 45)\}$, with $\mathcal{I} = \{12, 123\}$, we can have $1 \perp\!\!\!\perp 2 \mid (3, 4, 5)$ where X_4, X_5 are fixed to the reference category, a statement of much more limited scope

than the original one. Finally, in example 11, though $\mathcal{J} = \{125, 126\}$ *is smaller than before, with* $\mathcal{I} = \{12, 124\}$*, we can only impose* $1 \perp\!\!\!\perp 2 \mid (4, 5, 6)$ *with* X_5, X_6 *fixed to the reference category.*

As mentioned above, the conditionally independence in M would not be restricted if the elements of $\mathcal{H}$ did not belong to $\mathcal{A}$, however, Theorem 1 implies that the resulting model is non smooth.

Example 16. *In example 3 suppose that all variables have the same number of categories; because* $\mathcal{A} = \{12, 123, 124, 1234\}$*, the conditional independence is not affected if we take* $\mathcal{H} = \{23, 234\}$*; though this corresponds to the same number of parameters as* $\mathcal{I}$*, the jacobian* $\boldsymbol{J}$ *of the LM algorithm is the identity matrix and the model is non smooth.*

Acknowledgments

We are grateful to M. Drton for useful discussions.

Appendix

Interactions defined as contrasts of averages of logarithms of probabilities.

An alternative to the *reference category* interactions $\eta_{I;M}(\boldsymbol{x}_I)$ are the interactions based on contrasts of averages which may be defined as

$$\bar{\eta}_{I;M}(\boldsymbol{x}_I \mid \boldsymbol{x}_{M\setminus I}) = \sum_{b\subseteq I} \frac{1}{\langle I \setminus b\rangle}(-1)^{|I\setminus b|} \sum_{\boldsymbol{x}_{I\setminus b}} \log(\pi(\boldsymbol{x}_b, \boldsymbol{x}_{I\setminus b}, \boldsymbol{x}_{M\setminus I})), \tag{7}$$

where $\langle I\rangle$ denotes the number of possible configurations of the vector $\boldsymbol{x}_I$.

When $M \setminus I$ is not empty, the interactions defined in (7) depend on the value of the remaining variables and may be interpreted as the log-linear interaction I in the marginal distribution M conditionally on $X_h = x_h \, \forall h \in \mathrm{M\backslash I}$. To use the interactions defined in (7) as a parameterization, the usual way of removing redundancies is to average with respect the conditioning variables, leading to the following expression

$$\begin{aligned}
\bar{\eta}_{I;M}(\boldsymbol{x}_I) &= \frac{1}{\langle M \setminus I\rangle} \sum_{\boldsymbol{x}_{M\setminus I}} \bar{\eta}_{I;M}(\boldsymbol{x}_I \mid \boldsymbol{x}_{M\setminus I}) \\
&= \frac{1}{\langle M \setminus I\rangle} \sum_{\boldsymbol{x}_{M\setminus I}} \sum_{b\subseteq I} \frac{1}{\langle I \setminus b\rangle}(-1)^{|I\setminus b|} \sum_{\boldsymbol{x}_{I\setminus b}} \log(\pi(\boldsymbol{x}_b, \boldsymbol{x}_{I\setminus b}, \boldsymbol{x}_{M\setminus I})) \\
&= \frac{1}{\langle M \setminus b\rangle} \sum_{\boldsymbol{x}_{M\setminus b}} \sum_{b\subseteq I} (-1)^{|I\setminus b|} \log(\pi(\boldsymbol{x}_b, \boldsymbol{x}_{M\setminus b})).
\end{aligned}$$

It is well known that both the *contrasts of averages* interactions $\bar{\eta}_{a;M}(\boldsymbol{x}_a)$, used by Bergsma and Rudas [4], and the *reference category* interactions $\eta_{a;M}(\boldsymbol{x}_a) = \eta_{a;M}(\boldsymbol{x}_a \mid \mathbf{0}_{\mathrm{M\backslash a}})$, used in this paper, are a parametrization of the joint probabilities.

Let $\bar{\boldsymbol{\eta}}(I, M; \mathbf{0}_{M\setminus I})$ denote the vector of log-linear interactions in (7) when the variables in $\boldsymbol{x}_I$ take all possible configurations in lexicographic order. It is easy to verify that we may write $\bar{\boldsymbol{\eta}}(I, M; \mathbf{0}_{M\setminus I}) = \boldsymbol{S}(I, M) \log \boldsymbol{p}(M)$, where $\boldsymbol{S}(I, M) = \bigotimes_{i\in M} \boldsymbol{S}_i$ and $\boldsymbol{S}_i$ is equal to the matrix $\boldsymbol{I} - \mathbf{1}\mathbf{1}'/(r_i + 1)$ without the first row if $i \in I$ and to the vector $(1 \quad , \mathbf{0}')$ otherwise. It is also easy to verify that the vector of log-linear interactions $\bar{\boldsymbol{\eta}}(I, M)$, obtained by averaging over all possible configurations of the conditioning variables, may be written as $\bar{\boldsymbol{\eta}}(I, M) = \bar{\boldsymbol{S}}(I, M) \log \boldsymbol{p}(M)$, where $\bar{\boldsymbol{S}}(I, M) = \bigotimes_{i\in M} \bar{\boldsymbol{S}}_i$ and $\bar{\boldsymbol{S}}_i$ is equal to $\boldsymbol{S}_i$ when $i \in I$ and to the vector $\mathbf{1}'/(r_i + 1)$ otherwise.

Lemma 8.

$$\bar{\boldsymbol{\eta}}(I, M; \mathbf{0}_{M\setminus I}) = \boldsymbol{A}\boldsymbol{\eta}(I, M) \tag{8}$$

$$\bar{\boldsymbol{\eta}}(I, M) = \boldsymbol{B}\boldsymbol{\eta}(\mathcal{I}, M) \tag{9}$$

for suitable matrices of constants $\boldsymbol{A}$ *and* $\boldsymbol{B}$.

Proof. By substitution in (5), $\bar{\boldsymbol{\eta}}(I, M; \mathbf{0}_{M\setminus I}) = \boldsymbol{S}(I, M)\boldsymbol{G}(M)\boldsymbol{\eta}(M)$ and (8) follows by noting that the kronecker product contains a 0 factor if there is an $i \in I$, $i \notin J$, because $\boldsymbol{S}_i$ is a matrix of row contrasts and $\boldsymbol{G}_i$ is the unitary vector; the same result arise if there is a $i \notin I$, $i \in J$, because $\boldsymbol{S}_i$ is the vector $(1 \quad , 0')$ and $\boldsymbol{G}_i$ is the $\bar{\boldsymbol{I}}_i$ matrix whose first row is a row of 0's. Equation(9) follows by a similar argument: when $i \in I$, $i \notin J$, $\bar{\boldsymbol{S}}_i = \boldsymbol{S}_i$ and we get a 0 factor as above, instead, when $i \notin I$, $i \in J$, $\bar{\boldsymbol{S}}_i$ is proportional to the unitary vector so that $\bar{\boldsymbol{S}}_i\boldsymbol{G}_i$ is also proportional to a unitary vector. □

By noting that any category may be chosen as reference category for each variable, (8) implies that any log-linear interaction I defined in (7) is a linear function of the log-linear interactions I defined in (1) for all possible values of $\boldsymbol{x}_I$. Instead, the log-linear interactions $\bar{\boldsymbol{\eta}}(I, M)$, obtained by averaging across the conditioning variables, are a linear function of all $\boldsymbol{\eta}(J, M)$, $I \subseteq J \subseteq M$.

Proofs of the Lemmas

Proof. **Proof of Lemma 4**. Note that $\boldsymbol{A}$ is the residual variance in a linear model where the binary variables indexed by $\mathcal{H}$ are regressed on the variables in $\mathcal{R}$ while $\boldsymbol{B}$ is the residual variance when $\mathcal{H}$ is regressed on the variables in $\mathcal{R}$ and $\mathcal{I}$. Then, properties of linear projections imply that we may write $\boldsymbol{C} = \boldsymbol{A} - \boldsymbol{B}$ where

$$\boldsymbol{C} = \boldsymbol{Q}'_{\mathcal{I}\mathcal{H}|\mathcal{R}} \left(\boldsymbol{F}_{\mathcal{I}\mathcal{I}} - \boldsymbol{F}_{\mathcal{I}\mathcal{R}}\boldsymbol{F}^{-1}_{\mathcal{R}\mathcal{R}}\boldsymbol{F}_{\mathcal{R}\mathcal{I}}\right)^{-1} \boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$$

is clearly a positive semi-definite matrix. Then Theorem 7.7.3 in Horn [11] implies that the spectral radius of $\boldsymbol{B}\boldsymbol{A}^{-1}$, which is equal to that of $\boldsymbol{A}^{-1}\boldsymbol{B}$, is always not greater than 1. The spectral radius is exactly 1 if and only if $\boldsymbol{C}$ or. equivalently, $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$ is singular. □

Proof. **Proof of Lemma 5**. (i) Suppose there is a v : $v \in \mathcal{H}$, $v \notin \mathcal{A}$ and consider the intersections of v with the elements of $\mathcal{I} \cup \mathcal{R}$; these must belong to $\mathcal{V}$ and cannot be contained in $\mathcal{A}$, hence they must belong to $\mathcal{R}$; the argument in the proof of Theorem 2 implies that the corresponding columns in $\boldsymbol{Q}_{\mathcal{I}\mathcal{H}|\mathcal{R}}$ are 0. (ii) Under independence

$\boldsymbol{P}_a\boldsymbol{G}_v = \bigotimes_{j=1}^{d} \boldsymbol{\Pi}_j$, where the factors $\boldsymbol{\Pi}_j$ are the entries of the last row of Table 1, $a \in (\mathcal{I} \cup \mathcal{R})$ and $v \in \mathcal{H}$. If there is a variable $j \in v$ which is not contained in any element of $\mathcal{I} \cup \mathcal{R}$, the last entry in the forth column of Table 1 indicates that there will be a factor $\mathbf{1}_j \bar{\boldsymbol{\pi}}'_j$ where $\bar{\boldsymbol{\pi}}'_j$ has r_j columns; the result follows from an argument similar to the one at the beginning of the proof of Theorem 2. □

Table 1:

	$j \in a$			$j \notin a$		
	$j \in t$	$j \in h$	$j \notin (t \cup h)$	$j \in t$	$j \in h$	$j \notin (t \cup h)$
$\boldsymbol{P}_a(j)$	$\boldsymbol{I}_j$	$\boldsymbol{I}_j$	$\boldsymbol{I}_j$	$\mathbf{1}_j \boldsymbol{\pi}'_j$	$\mathbf{1}_j \boldsymbol{\pi}'_j$	$\mathbf{1}_j \boldsymbol{\pi}'_j$
$\boldsymbol{G}_{t,h}(j)$	$\bar{\boldsymbol{I}}_j$	$\boldsymbol{e}_{jl}$	$\mathbf{1}_j$	$\bar{\boldsymbol{I}}_j$	$\boldsymbol{e}_{jl}$	$\mathbf{1}_j$
Π_j	$\bar{\boldsymbol{I}}_j$	$\boldsymbol{e}_{jl}$	$\mathbf{1}_j$	$\mathbf{1}_j \bar{\boldsymbol{\pi}}'_j$	$\mathbf{1}_j \pi_{jh}$	$\mathbf{1}_j$

Proof. **Proof of Lemma 6**. When $\mathcal{K}(t)$ is not a singleton, by construction, the intersection of two or more elements of $\mathcal{K}(t)$ is disjoint from h, thus $\mathcal{K}(t, h)$ is formed by sets $\mathcal{G}$ with cardinality not smaller than two. Let n_t is the cardinality of $\mathcal{K}(t)$ then:

$$\sum_{\mathcal{G} \in \mathcal{K}(t,h)} (-1)^{|\mathcal{G}|+1} = \sum_{i=2}^{n_t} \left(\begin{array}{c} n_t \\ i \end{array} \right) (-1)^{i+1} = -\sum_{i=0}^{1} \left(\begin{array}{c} n_t \\ i \end{array} \right) (-1)^{i+1} = -n_t + 1,$$

thus, point (ii) of Definition 3 is satisfied. Point (iii) is trivially satisfied because $\mathcal{I}$ is a singleton. When $\mathcal{K}(t)$ is a singleton all the conditions of Definition 1 are trivially satisfied. □

Proof. **Proof of Lemma 7**. Under independence $\boldsymbol{P}_a\boldsymbol{G}_{t,h} = \bigotimes_{j=1}^{d} \boldsymbol{\Pi}_j$, where $\boldsymbol{\Pi}_j = \boldsymbol{P}_a(j)\boldsymbol{G}_{t,h}(j)$; the possible values of $\boldsymbol{\Pi}_j$ are given in Table 1 where $\bar{\boldsymbol{I}}_j$ is an identity matrix without the first column, $\mathbf{1}_j$ is a vector of ones $\boldsymbol{e}_{jl}$ a vector of 0's except for a 1 in the lth position, and $\boldsymbol{\pi}_j$ is the marginal distribution of X_j, all of dimension $r_j + 1$. Point a) follows from the first two columns of Table 1, while b) follows from columns 1 and 5. □

References

[1] Agarwal, R.P., M. M., ORegan, D., 2001. Fixed point theory and applications. Cambridge Univeristy Press.

[2] Barndorff-Nielsen, O. E., 1979. Information and exponential families in statistical theory. Wiley and Sons, New York.

[3] Bartolucci, F., Colombi, R., Forcina, A., 2007. An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. Statist. Sinica 17 (2), 691.

[4] Bergsma, W. P., Rudas, T., 2002. Marginal models for categorical data. Ann. Statist. 30 (1), 140–159.

[5] Drton, M., 2009. Discrete chain graph models. Bernoulli 15, 736–753.

[6] Drton, M., Xiao, H., 2010. Smoothness of gaussian conditional independence models. Contemporary Mathematics 518, 155–177.

[7] Evans, R. J., 2011. Smoothness of discrete conditional independence models, personal comunication.

[8] Forcina, A., 2011. Smoothness of conditional independence models for discrete data, submitted.

[9] Forcina, A., Lupparelli, M., Marchetti, M., 2010. Marginal parameterizations of discrete models defined by a set of conditional independencies. Journ. Mult. Analysis 101 (10), 2519–2527.

[10] Hojsgaard, S., 2004. Statistical inference in context specific interaction models for contingency tables. Scandinavian Journal of Statistics 31, 143–158.

[11] Horn, R.A., J. C., 2009. Matrix analysis. Cambridge Univeristy Press.

[12] Roverato, A., Lupparelli, M., La Rocca, L., 2012. Log-mean models for binary data. arXiv.org 1109.6239, 1–36.

[13] Rudas, T., Bergsma, W. P., Németh, R., 2010. Marginal log-linear parameterization of conditional independence models. Biometrika 97 (4), 1006–1012.